

Міністерство освіти і науки України Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

КРИПТОГРАФІЯ

Комп'ютерний практикум №1

«Експериментальна оцінка ентропії
на символ джерела відкритого тексту»

Виконав

студент групи ФБ-93

Флекевчук Данило

Перевірила

Селюх П.В.

Київ – 2021

Мета:

Засвоїти поняття ентропії та надлишковості мови. Навчитись проводити частотний аналіз тексту та за його результатами рахувати ентропію тексту. Покращити навички програмування.

Завдання:

- уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- написати програму для підрахунку частот літер і частот біграм в тексті, а також підрахунку H_1 , H_2 за безпосереднім означенням. Підраховувати частоти літер та біграм, а також значення H_1 , H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1 Мб), де ймовірності замінюються відповідними частотами. Також отримати значення H_1 , H_2 на тому ж самому тексті, в якому вилучено всі пробіли.
- за допомогою програми CoolPinkProgram оцінити значення $H(10)$, $H(20)$, $H(30)$.
- використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

I.

В цій частині роботи я рахую частоту входження літер, біграм(двох типів) у текстовий файл, та за цими показниками рахуємо ентропію мови. Для обчислення було використано мову програмування Python, так як вона найкраще підходить для важких обчислень. Для підрахунку входжень я використовував модуль collections а саме Counter для підрахунку входжень у текст та OrderdDict для сортування. Я розбив текст на різні типи біграм, з допомогою циклів з різними кроками (перехрестні: 2, прості: 1).

Код знаходиться в тій же папці що і протокол.

Далі я порахував ентропію для біграм та літер, і посортував літери за зростанням частоти. І вивів розраховані значення ентропії в консоль. Ентропію я рахував за формулою, що була наведена в завданні.

```
PS C:\Users\Данил\Desktop\CriptaLab1> python main.py
ЛІТЕРИ
ЕНТРОПІЯ: 4.449600333417484
о : 96681, е : 73380, а : 67116, н : 54835, и : 54644, т : 54553, с : 44600, в : 38984, л : 38734, р : 35246, к : 27833, д : 26983, м : 26495, у : 24993, п : 23121, ь : 19566, я : 18000, ч : 15249, б : 14655, г : 14232, ы : 13912, э : 12975, ж : 9615, й : 8437, х : 7171, ш : 6937, ю : 4733, э : 2971, щ : 2521, ц : 2336, ф : 1049, ё : 65,

ПРОСТІ БІГРАМИ
ЕНТРОПІЯ: 8.251331417953548

ПЕРЕХРЕСТНІ БІГРАМИ
ЕНТРОПІЯ: 8.254367201998619
PS C:\Users\Данил\Desktop\CriptaLab1> █
```

Для матриць з біграмами було виділено два файли, test_gap.txt та test.txt. Туди я повиводив значення частот для біграм у формі матриць, де на перехресті літер стоять біграми.

Без пробілів

ПРОСТІ БІГРАМИ

[illegible][illegible]

and

[illegible]

```
я 0.011327 0.000000
```

А Б В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Э Ю Я																																
0.003606	0.002508	0.006311	0.016676	0.002527	0.000806	0.004101	0.000000	0.002243	0.004399	0.011555	0.000001	0.008856	0.002414	0.005756	0.015874	0.011219	0.015842	0.004816	0.015584	0.008765	0.004562	0.003935	0.001545	0.000288	0.006128	0.000563	0.000666	0.000000	0.000001	0.002375	0.000043	0.006259
0.016807	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000438	0.000012	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.005597	0.005649	0.000002	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.001028	0.000998	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.018882	0.000013	0.001233	0.001890	0.003280	0.002683	0.001734	0.000000	0.000787	0.001180	0.000186	0.002110	0.000148	0.005524	0.000430	0.006458	0.000255	0.001136	0.005381	0.004321	0.005670	0.000156	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.001151	0.004447	0.000115	0.000563	0.000342	0.000556	0.000373	0.000000	0.000406	0.000212	0.000348	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.017332	0.000000	0.000056	0.002430	0.000237	0.001700	0.001655	0.000000	0.000000	0.000000	0.000000	0.00																					

Произвольная часть текста:
ы_о_нем_аналогично_

Использованные буквы:

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ:

Символ по счету:

Номер эксперимента: 51

Неравенство для энтропии:
 $2,52628436774245 < H < 3,13669456263237$

Двоичная таблица угаданных символов:

Поле ввода символов:

Продолжить Другой

Вероятности:

q[1] = 0,42
q[2] = 0,1
q[3] = 0,08
q[4] = 0
q[5] = 0,02
q[6] = 0,02
q[7] = 0,06
q[8] = 0
q[9] = 0,02
q[10] = 0,02
q[11] = 0
q[12] = 0,02
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0
q[17] = 0,04
q[18] = 0,02
q[19] = 0
q[20] = 0
q[21] = 0,04
q[22] = 0
q[23] = 0,02
q[24] = 0
q[25] = 0,06
q[26] = 0
q[27] = 0
q[28] = 0,02
q[29] = 0
q[30] = 0,02
q[31] = 0,02
q[32] = 0

Строка состояния:

Произвольная часть текста:
еши_как_бы_вам_понравилось_ес

Использованные буквы:

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ:

Символ по счету:

Номер эксперимента: 52

Неравенство для энтропии:
 $1,64952187862109 < H < 2,24337732513425$

Двоичная таблица угаданных символов:

Поле ввода символов:

Продолжить Другой

Вероятности:

q[1] = 0,5686274
q[2] = 0,1568627
q[3] = 0,0784313
q[4] = 0
q[5] = 0
q[6] = 0
q[7] = 0
q[8] = 0
q[9] = 0
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0,019607
q[14] = 0,019607
q[15] = 0
q[16] = 0,019607
q[17] = 0
q[18] = 0,019607
q[19] = 0,039215
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0,019607
q[25] = 0,019607
q[26] = 0
q[27] = 0,019607
q[28] = 0
q[29] = 0
q[30] = 0,019607
q[31] = 0
q[32] = 0

Строка состояния:

$H(10) = 3$; $H(20) = 2,8314$; $H(30) = 1,9464$

Для обрахунку ентропії, я просумував усі значення ентропій і поділив їх на 3.

$H = (3 + 2,8314 + 1,9464)/3 = 2,5926$

Ентропія	Значення ентропії	Надлишковість
$H(10)$	3.0000	0.4000
$H(20)$	2.8140	0.4372
$H(30)$	1.9460	0.6108
H	2,5926	0.4800

Надлишковість рахується за формулою $1 - (H/H(0))$, де $H(0)$ для російської мови рівне 5. У мене це вийшло 0,48, я думаю такт мала надлишковість зумовлена моїми неякісними знаннями російської мови.

Висновок:

В ході роботи я навчився обраховувати ентропію джерела тексту для символів та біграм, також я навився обраховувати надлишковість тексту. Для російської мови надлишковість склала 48%, за вказаними вище причинами. Для пошуку ентропії біграм та літер було обрано “Злочин і кара”. Якщо не забрати пробіли то, тоді вони будуть найбільш вживаним символом (“_” : 172811), якщо ні то це літери: “о” : 96681, “е” : 73380 та “а” : 67116 відповідно, а най менш вживані: “ц” : 2336, “ф” : 1049. Найчастішими біграмами з пробілом виявились: “о_” : 0.023578, “е_” : 0.018882, “и_” : 0.017332, а без пробілів: “то” : 0.018084, “ов” : 0.012589, “не” : 0.012239. Наявність пробілів дуже полегшує роботу криптоаналітика.